

Conference Abstract

Validation for preservation: how sustainable are GBIF datasets?

Joakim Philipson ‡

‡ Stockholm University, Stockholm, Sweden

Corresponding author: Joakim Philipson (jomtov@yahoo.com)

Received: 13 Apr 2018 | Published: 22 May 2018

Citation: Philipson J (2018) Validation for preservation: how sustainable are GBIF datasets? Biodiversity Information Science and Standards 2: e25805. <https://doi.org/10.3897/biss.2.25805>

Abstract

Validation using schemas and tools like the Darwin Core Archive Validator from GBIF are mainly seen as methods of checking data quality and fitness for use, but are also important for long-term preservation. We may like to think that our present (meta)data standards and formats are made for eternity, but in reality we know that standards evolve, formats change (some even become obsolete with time), and so do our needs for storage, searching and future dissemination for re-use. So we might eventually come to a point where transformation of our archival records and migration to other formats will be necessary. This could also mean that even if the AIPs, the Archival Information Packages stay the same in storage, the DIPs, the Dissemination Information Packages that we want to extract from the archive are subject to change of format. Further, in order for archival information packages to be self-sustainable as required in the OAIS model, it is important to take interdependencies between individual files in the information packages into account, already by the time of ingest and validation of the SIPs, the Submission Information Packages, and along the line at different points of necessary transformation / migration (from SIP to AIP, from AIP to DIP etc.) to counter obsolescence. Validation schemas and transformation code should also be archived together with the AIPs. By ensuring compliance with standards these tools are essential in controlling *uniformity* of records in a collection, for future needs of transformation and migration to new, sustainable formats. An example is given of the problems encountered in transforming only a small, relatively well

defined collection of about 1000 archival items but with substantial variations between them, due to a lack of effective input constraints and validation at ingest.

A further assessment is made of validation errors encountered in some Darwin Core Archives comprising thousands of records from some hundred published datasets, and how these errors might affect a future potential transformation / migration effort. Migration efforts must necessarily be general in scope, while errors in datasets from non-compliance with standards risk being reinforced or aggravated in the transformation process, making the information contained in the resulting records more difficult to interpret. The conclusion is that efforts should be made, e.g. by means of embedded validation measures into upload forms and other methods of information transfer (e.g. ftp, oai-pmh) to ensure as close compliance as possible to standards, already at the time of ingest.

Keywords

validation, preservation, migration, transformation, GBIF, Darwin Core Archive

Presenting author

Joakim Philipson <https://orcid.org/0000-0001-5699-994X>

Presented at

TDWG 2018

Hosting institution

Stockholm University Library